LinguIUTics at PsyDefDetect: Iterative Imbalance-Aware Fine-tuning of Qwen3-8B for Psychological Defense Mechanism Classification

Shefayat E Shams Adib*, **Ahmed Alfey Sani***, **Md Hasibur Rahman Alif***, **Ajwad Abrar***

Department of Computer Science and Engineering,
Islamic University of Technology, Dhaka, Bangladesh
{shefayatadib, ahmedalfey, hasiburrahman21, ajwadabrar}@iut-dhaka.edu

*All authors contributed equally to this work.

Abstract

Detecting psychological defense mechanisms in conversational text remains a challenging clinical NLP problem. For the PsyDefDetect 2026 shared task (9-class utterance classification evaluated via macro F1), our team LinguIUTics[1] achieves a macro F1-score of 0.3917 on the official positive-class leaderboard, ranking 4th out of 21 registered teams and improving over the Ministral-8B task baseline (31.48 macro F1) by +7.7 absolute points (+24.4% relative). BERT-family encoders and zero-shot LLMs proved ineffective on rare classes due to severe class imbalance, leading us to QLoRA fine-tuning of Qwen3-8B. We leverage three key strategies: grouped stratified cross-validation (preventing leakage), minority-class round-robin lexical augmentation, and a post-processing pipeline with logit bias tuning and ensemble blending. Together, these components close much of the validation–leaderboard gap and substantially improve minority-class recall, driving the critical "Unclear" class (Level 8) from near-zero performance to F1 = 0.797.

1 Introduction

Automatic detection of psychological defense mechanisms (unconscious strategies to mitigate distress under the DMRS framework (Perry and Henry, 2004)) helps mental health platforms flag maladaptive coping and improves empathetic conversational agents (Liu et al., 2021; Na et al., 2025). The PsyDefDetect 2026 shared task (Na et al., 2026a,b) challenges participating systems to classify seeker utterances into nine DMRS levels. This poses a major obstacle, that is extreme class imbalance (He and Garcia, 2009), with the frequency gap between majority ("High-Adaptive", 51.9%) to minority ("Unclear ",1.5%) classes at about 34.6 times respectively. Because evaluation uses macro-averaged F1, optimizing on accuracy leads to majority-class collapse and task failure.

To address this, we followed up with an iterative development process. Standard single-fold PEFT (parameter-efficient fine-tuning) on Qwen3-8B (Yang et al., 2025) suffered a massive generalization gap (0.34 validation vs. 0.24 leaderboard F1) due to limited low-rank capacity and majority-class overfitting. By systematically upgrading model capacity, loss functions, and inference, we established a robust pipeline. Our key contributions are:

1. A leakage-safe cross-validation scheme at the level of groups, where synthetic augmentations are in a set with their source utterances. This leads to an order of magnitude smaller generalization gap between out-of-fold and leaderboard batches.

2. An oversampling method that preserves the original psychological signal(Wei and Zou, 2019). This is achieved by expanding specific minority classes by 3 times in a round-robin lexical mutation approach.

3. A post-processing pipeline that combines OOF based logit bias tuning, that is guarded using v2 decoding, and multi-seed probability blending.

2 Task and Dataset

The PsyDefDetect 2026 task (Na et al., 2026a) classifies seeker utterances into nine psychological defense levels, as defined by the DMRS framework (Perry and Henry, 2004), and evaluated by macro averaged F1-score. The PsyDefConv dataset (Na et al., 2026b) contains 2,336 utterances across 200 dialogues from ESConv (Liu et al., 2021). For 5-fold CV, we merged train and validation sets (1,864 training examples and 472 test). It exhibits very

[1]Code and resources are available at https://github.com/Shefwef/LingIUTics-PsyDefDetect-BIONLP26

Proceedings of the BioNLP 2026 (Shared Tasks), pages 123–131
July 3-4, 2026 ©2026 Association for Computational Linguistics

high class imbalance with a 34.6 times frequency gap (Table 1).

L	Defense Mechanism	N	%
0	No Defense / Neutral	296	15.9
1	Action Defenses	108	5.8
2	Major Image-Distorting	61	3.3
3	Disavowal	99	5.3
4	Minor Image-Distorting	84	4.5
5	Neurotic	48	2.6
6	Obsessional	172	9.2
7	**High-Adaptive**	**968**	**51.9**
8	Unclear / Needs More Info	28	1.5
	Combined Train	**1,864**	**100.0**

Table 1: PsyDefConv combined training class distribution (Train + Val splits). Level 7 vs. Level 8 ratio: 34.6×.

3 Implementation Process

Our development went through three iterative stages, each revealing a fundamental limitation that directly inspired the next architectural transition. Table 10 follows the complete leaderboard path from R0 (F1 = 0.240) to our final submission (F1 = 0.392). The detailed system run log is provided in Table 2.

System / Variant	OOF F1	LB F1
R0: MentalBERT	N/A	0.240
R1: MentalBERT+RoBERTa ensemble	N/A	0.240
R2: MentalRoBERTa Focal + EMD	0.314	N/A
R3: DeBERTa-v3-base 5-fold	0.307	0.236
R4: RoBERTa-base	N/A	0.269
R5: Qwen3-8B 1-fold r=64 baseline	0.345	0.249
R6: Qwen3-8B 5-fold r=128 + weighted CE	0.361	0.329
R7: Qwen3-8B v2	0.372	0.355
R8: Qwen3-8B v2 microplus	0.372	0.354
R9: Qwen3-8B seed A only + v2 decode	0.431	N/A
R10: Qwen3-8B old + seed A blend + v2 decode	**0.437**	**0.392**

Table 2: Complete system run log. LB = CodaBench leaderboard.

3.1 Stage 1: BERT-Family Encoder Baselines

We evaluated MentalBERT, MentalRoBERTa, DeBERTa-v3-base, and RoBERTa-base (Devlin et al., 2019; Ji et al., 2022; He et al., 2021; Liu et al., 2019) for multiple context windows, loss functions (cross-entropy, Focal (Lin et al., 2017), EMD), and ensemble strategies (full scores in Appendix E, Table 10). Best validation macro F1 was 0.314 (MentalRoBERTa + Focal/EMD + Hungarian remapping), with leaderboard peak at 0.240. Importantly, F1 for Classes 3, 5 and 8 was still zero across *all* variants, establishing an *encoder capacity ceiling* with a 51.8% majority-class prior using $n \leq 50$ minority examples. From the general formula:

$$\mathcal{L}_{\text{Focal}} = -(1 - p_t)^{\gamma} \log p_t \quad (1)$$

where p_t is the probability of predicting the correct class, logarithmically compensated by strength constant γ.

3.2 Stage 2: Zero-Shot Evaluation

Qwen3-8B, Llama 3.1-8B, and Ministral-8B evaluated zero-shot with explicit DMRS label definitions produced 8–16% macro F1 (Table 10), confirming that task knowledge cannot be prompt-engineered.

3.3 Stage 3: Diagnostic LLM Fine-Tuning

Ministral-8B fine-tuned with 4-bit NF4 quantization achieved 64.71% accuracy but only 14.74 macro F1 (Table 10). This illustrates that standard cross-entropy collapses to majority-class prediction under severe imbalance. Furthermore, the accuracy is an actively misleading metric in this setting.

3.4 Stage 4: Final System (Qwen3-8B LoRA Pipeline)

As such, with the three lessons above guiding us, our final pipeline consists of five components: model architecture, imbalance-aware training objective, data augmentation, leakage-safe cross-validation and a post-processing ensemble, each targeting a specific failure mode identified in the earlier stages.

3.4.1 Model Architecture

We fine-tune Qwen3-8B via QLoRA (Dettmers et al., 2023) with 4-bit NF4 quantization, reducing peak GPU memory from ∼32 GB to ∼8 GB on a single NVIDIA RTX 3090 Ti. LoRA adapters (Hu et al., 2021) target all attention and feed-forward layers (`q`, `k`, `v`, `o`, `gate`, `up`, `down`) and the `score` head ($r = 128$, $\alpha = 256$, dropout = 0.1), yielding ≈31M trainable parameters (0.4% of the 8B base). Increasing rank from $r = 64$ to $r = 128$ delivered a +24.9% fold-level F1 gain, critical for separating psychologically similar classes (e.g., Level 4 vs. Level 5). Full training hyperparameters for this configuration are summarised in Table 8 in Appendix B.

3.4.2 Input Representation

Each prompt has three parts: (1) the DMRS Label Guide with 9-class clinical schema; (2) Conversational Context of the last 30 dialogue turns

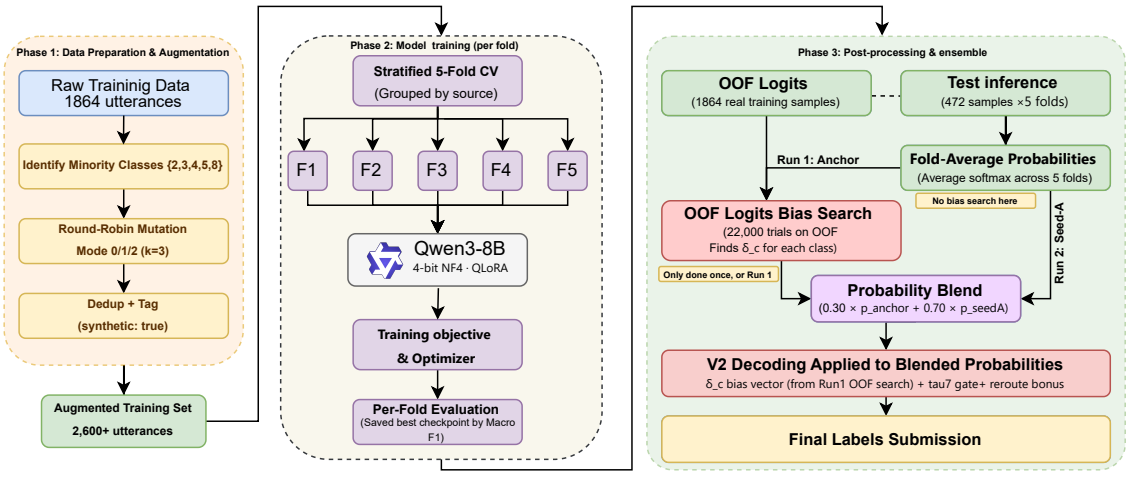


Figure 1: Full system pipeline. **Phase 1**: raw training data is preprocessed and minority classes (Levels 2, 3, 4, 5, 8) are oversampled via round-robin lexical mutation, creating an expanded training set of 2,600+ utterances. **Phase 2**: two independent grouped stratified 5-fold QLoRA fine-tuning runs of Qwen3-8B the *Anchor* (seed = 42) and *Seed-A* (seed = 20260407) sharing identical architecture and hyperparameters, with dialogues grouped across folds to prevent data leakage. **Phase 3**: a class-specific logit bias vector (δ_c) is optimised on Anchor OOF predictions and locked for reuse; test probabilities from both runs are blended before a final guarded decode step that maximises minority-class recall without sacrificing majority-class precision.

prefixed by `SEEKER:`/`HELPER:` tags and (3) an Output Instruction directing the model to emit a single integer (0–8). Inputs are tokenised to 1,024 tokens with dynamic padding to multiples of 8, covering >95% of samples without truncation.

3.4.3 Imbalance-Aware Training Objective

This dataset is, to a large extent, imbalanced towards Level 7. To mitigate the problem of majority-class collapse, we apply two stabilization techniques.

Inverse-Square-Root Class Weighting. Per-class weights

$$w_c = \frac{1/\sqrt{n_c}}{\sum_{i=0}^{8} 1/\sqrt{n_i}} \qquad (2)$$

boost the most under-represented classes (e.g., $w_8 = 1.67$, $w_5 = 1.29$) while dampening the majority class ($w_7 = 0.28$), instead of inverse-frequency weighting that can lead to gradient instabilities.

Label Smoothing & Optimization Schedule. By using label smoothing (Szegedy et al., 2016) ($\varepsilon = 0.05$), we avoid early logit saturation for Level 7. This prevents gradients from being dominated by the majority class and allows minority-class signals to contribute more effectively during back-propagation. We employ AdamW with a peak learning rate of 1.2×10^{-4}, cosine annealing (8% warmup), 10 epochs per fold, batch size 16 (2×8 accumulation), and gradient clipping at 0.3 (see Appendix B).

3.4.4 Data Augmentation

For the rarest classes (Levels 2, 3, 4, 5, and 8), there are between 28 and 84 training examples. This is not enough for an 8B capacity model to learn sufficiently reliable decision boundaries. In order to fix this, we perform **round-robin lexical mutation** to generate $k = 3$ synthetic variants per source utterance in these classes, cycling through three surface-level rewriting modes:

- **Mode 0:** Contraction replacement (e.g., *I am* → *I'm*) plus a hedging prefix (*Honestly, . . .*).
- **Mode 1:** Vocabulary style-shift (e.g., *maybe* → *perhaps*) plus a trailing filler (*. . . I guess.*).
- **Mode 2:** Hesitation markers (e.g., periods → ellipses; ? → ??).

Mutations target only the seeker utterance to preserve the psychological signal; after deduplication, minority class counts increased from 28–84 to 65–252 examples (see Appendix C). This targeted minority-class oversampling strategy is consistent with prior findings that augmenting only underrepresented classes yields more effective and stable performance improvements than augmenting all classes equally (Sani et al., 2026).

3.4.5 Grouped Five-Fold Cross-Validation

Random splitting risks leakage across the 200 source dialogues, making dialogue-level isolation essential. We therefore apply grouped stratified cross-validation (grouped CV, implemented as `StratifiedGroupKFold` with $k = 5$) using `dialogue_id` as the grouping key:

- **Zero Leakage Guarantee** (0 leaked dialogues confirmed): All utterances and their synthetic variants are kept entirely within one fold.
- **Reliable Validation Signal** (OOF leaderboard gap reduced from 9.6 to 1.7-4.5 points): Strong rank-correlation between OOF and leaderboard gains enables safe threshold tuning.
- **Ensemble Foundation** (5 checkpoints per seed): Five-fold training yields pure OOF predictions for post-processing calibration and reduces inference variance.

Reliable validation behaviour across folds is shown in the per-fold OOF metrics in Table 3.

Fold	Acc.	Mac. F1	Mac. P	Mac. R
1	0.6193	0.3804	0.3836	0.3925
2	0.6247	0.3701	0.3902	0.3617
3	0.6408	0.3899	0.4276	0.4000
4	0.6300	0.3553	0.3753	0.3514
5	0.5968	0.3326	0.3396	0.3298
OOF	**0.6223**	**0.3716**	**0.3817**	**0.3675**

Table 3: Per-fold CV results (grouped-clean augmented run, before seed blending).

3.4.6 Post-Processing and Ensemble Strategy

Despite class-weighted training, raw probabilities remain heavily biased towards the majority class (Level 7). To rectify this and recover rare classes without compromising precision, we implement a three-stage post-training pipeline (`v2decode`).

Stage A: OOF Bias Optimization. Using logit adjustment for long-tail learning (Menon et al., 2021), we search for class-specific probability offsets (δ_c) that maximize the OOF macro F1 score:

$$\hat{y} = \arg\max_c \left[\log p_c + \delta_c\right] \tag{3}$$

We evaluate approximately 22,000 randomly sampled bias vectors on OOF predictions to identify a configuration that balances majority precision with minority recall. The best locked vector applies a negative penalty to Level 7 ($\delta_7 < 0$) and substantial positive bonuses to minority classes like Level 8 ($\delta_8 > 0$).

Stage B: Multi-Seed Blending. We run a second identical 5-fold training pipeline, denoted *Seed-A*, using a different random seed (`seed = 20260407`) and the same architecture and hyperparameters. We combine the test-set probabilities of the original Anchor and Seed-A using a 30/70 weighted average:

$$p_{\text{blend}} = 0.30 \cdot p_{\text{anchor}} + 0.70 \cdot p_{\text{seedA}} \tag{4}$$

The ratio was tuned using real-only OOF F1, combining the Anchor's high precision with Seed-A's strong minority recall.

Stage C: The τ_7-Gate Decoding. A confidence safeguard is applied from the locked δ_c bias vector onto p_{blend} to prevent precision collapse:

- **τ_7-Protection Gate:** The prediction is locked to Level 7 and δ_c offsets are not applied if $p_{\text{blend},7} \geq 0.69$.
- **Minority Rerouting:** If $p_{\text{blend},7} < 0.69$, δ_c offsets are applied, rerouting ambiguous samples into the highest-probability minority class.

So, minority labels are aggressively recovered when the model is uncertain. This increases the minority recall without affecting the precision.

4 Experiments and Results

4.1 Cross-Paradigm Comparison

BERT-family encoders struggled to break above 0.314 macro F1 due to capacity limits, across three paradigms (Table 10 in Appendix E). Both zero-shot LLMs and standard LLM fine-tuning collapsed to majority-class predictions (near 15% F1). In contrast, our imbalance-aware Qwen3-8B pipeline resolved these issues, reaching 39.17% macro F1.

4.2 Comparison with SOTA Baselines

Table 4 compares the results of our systems against the task baselines (Na et al., 2026a,b). Our final pipeline surpassed the stated SOTA, Ministral-8B fine-tuned baseline (31.48 macro F1) +7.7 absolute points, corresponding to a +24.4% relative improvement in macro F1.

System	Acc. (%)	Macro F1 (%)
GPT-5 zero-shot (task paper)	52.75	19.53
Gemini 2.5 Pro zero-shot	56.36	25.99
DeepSeek-V3.2 zero-shot (CoT)	55.72	26.17
Llama 3.1-8B fine-tuned	62.92	30.51
InternLM3-8B fine-tuned	63.98	30.53
Ministral-8B fine-tuned (SOTA)	64.83	31.48
DeBERTa-v3-base (5-fold)	59.11	23.58
RoBERTa-base	51.27	26.97
Qwen3-8B LoRA (Baseline Finetuned)	54.45	24.91
Qwen3-8B LoRA (Grouped CV + Bias Tuning)	58.43	35.48
Qwen3-8B LoRA (SeedA Ensemble + v2decode)	**64.19**	**39.17**

Table 4: Comparison with task paper baselines.

4.3 Ablation Study

Table 5 analyses each component's contribution. Increasing LoRA rank to $r = 128$ produced the highest boost (+24.9% fold-level F1), supporting

System Configuration	Macro F1
R0: 1-fold, r=64, no weighting	0.249
+ 5-fold CV, r=128	0.284†
+ Weighted CE + label smoothing	0.329†
+ Grouped-clean 5-fold	0.355
+ Data augmentation (RR-k3)	0.355
+ Seed-A blend (30/70) + v2 decode	**0.392**

Table 5: Ablation: each component's contribution. †Metrics for these rows are single-fold estimates from the 5-fold setup, included as indicative rather than full OOF results.

that model capacity was indeed the primary bottleneck. Grouped CV, data augmentation, and post-processing decode rules contributed incrementally, securing the final +3.69 F1 points.

4.4 Per-Class Analysis

Per-class performance is shown in Figure 2 (Appendix D) and Table 6. Level 8 ("Unclear") saw the most improvement, climbing from near-zero recall to 0.797 F1 via augmentation and bias tuning, with Levels 2 and 3 also gaining substantially. On the other hand, Levels 4 and 5 continue to be difficult (0.25–0.27 F1) due to high linguistic overlap with the majority class. Importantly, optimising for minority classes did not compromise the majority class (Level 7), which still resulted in a solid F1 of 0.709.

5 Conclusion

We showed that the data-centric imbalance mitigation methods (grouped CV, weighted loss, round-robin lexical augmentation, and dynamic OOF bias tuning with ensembling) that we used were much more important than raw model capacity for psychological defense classification. We achieved a macro F1-score of 0.3917 on the official positive-class leaderboard, ranking 4th out of 21 registered teams. This corresponds to a +7.7 macro F1-score improvement (+24.4% relative) over the Ministral-8B fine-tuned baseline. In future, we plan to add more effective paraphrase-based data augmentation, use losses better suited to imbalanced classes, and evaluate on more datasets.

L	Mechanism	P	R	F1
0	Neutral	0.747	0.858	0.799
1	Action	0.242	0.398	0.301
2	Major Img-D	0.480	0.451	0.465
3	Disavowal	0.401	0.402	0.401
4	Minor Img-D	0.317	0.211	0.254
5	Neurotic	0.398	0.214	0.278
6	Obsessional	0.203	0.267	0.231
7	High-Adaptive	0.693	0.726	0.709
8	Unclear	0.797	0.797	0.797
Macro		**0.431**	**0.436**	**0.426**

Table 6: Per-label OOF metrics, final blended system with v2 decode. Level 8 improved from ≈0 to 0.797 via augmentation and bias tuning.

Limitations

These decode rules and OOF bias vectors are calibrated specifically to this dataset and so requires recomputation for unseen domains. The grouped CV protocol keeps mutant variants within the same source group to reduce leakage, but the risk of leakage cannot be fully eliminated. Lastly, we were limited to PEFT on models with 8B parameters or less and by hardware constraints (24 GB VRAM).

Acknowledgments

We thank the PsyDefDetect 2026 shared task organizers (Na et al., 2026a) for providing the PsyDefConv dataset (Na et al., 2026b) and evaluation infrastructure through CodaBench.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, Venice, Italy.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

John Perry and Melissa Henry. 2004. Studying defense mechanisms in psychotherapy using the defense mechanism rating scales. *Advances in Psychology*, 136.

Ahmed Alfey Sani, Kazi Akib Zaoad, Shefayat E Shams Adib, Md Abdul Muqtadir, and Ajwad Abrar. 2026. Addressing data scarcity in bangla fake news detection: An llm-based dataset augmentation approach. *arXiv preprint arXiv:2605.01292*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, Nevada.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Full DMRS Label Definitions

Table 7 provides the complete clinical descriptions for all nine psychological defense levels used in our classification prompt.

B Full Hyperparameter Table

Hyperparameter	Value
Base model	Qwen/Qwen3-8B
Quantization	4-bit NF4 + double quant
LoRA rank / alpha	128 / 256
LoRA dropout	0.1
LoRA target modules	q/k/v/o/gate/up/down/score
Trainable parameters	≈31M (0.4%)
Max sequence length	1024
Optimizer	AdamW
Learning rate	1.2×10^{-4}
Weight decay	0.01
LR scheduler	Cosine annealing
Warmup ratio	8%
Per-device batch size	2
Gradient accumulation	8 (eff. batch = 16)
Gradient clip norm	0.3
Epochs per fold	10
Label smoothing ε	0.05
Class weight formula	Inverse-sqrt (Eq. 2)
Hardware	NVIDIA RTX 3090 Ti 24 GB
Mixed precision	bf16

Table 8: Complete training hyperparameters for the final Qwen3-8B LoRA system.

C Data Augmentation Examples

Below are three mutation modes applied to a single Level-3 (Disavowal) sample. All mutations target only the seeker utterance; the supporting dialogue context is unchanged.

D Confusion Matrices

Figure 2 illustrates error distributions across minority and majority classes, highlighting grouping and filtering improvements.

E Comprehensive Model Comparison

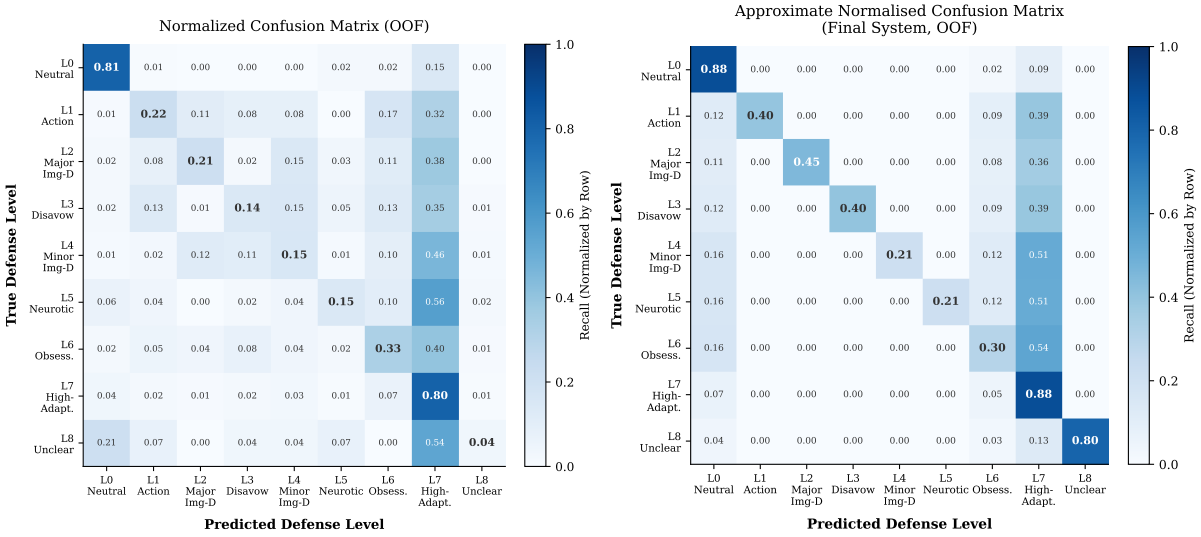
Table 10 reports all systems evaluated during our development, organised by model family and experimental stage. Empty cells indicate the model was not evaluated under that paradigm. Rows marked † are external baselines from the task paper (Na et al., 2026b); all others are our own internal tuning experiments.

L	Name	Clinical Description
0	No Defense	Phatic or factual exchange with no active defense mechanism
1	Action	Emotional discharge through behavior (passive aggression, complaining, impulsive action)
2	Major Img-D	Extreme cognitive distortion: projection, splitting, all-or-nothing thinking
3	Disavowal	Avoiding unpleasant reality: denial, rationalization, minimization
4	Minor Img-D	Subtler distortion: devaluation, idealization, omnipotence
5	Neurotic	Unconscious anxiety management: repression, displacement, reaction formation
6	Obsessional	Over-intellectualization, isolation of affect, undoing
7	High-Adaptive	Mature coping: humor, altruism, insight, self-assertion, sublimation
8	Unclear	Insufficient context for reliable DMRS classification

Table 7: DMRS label definitions used in the classification prompt.

Mode	Utterance Text
Original (Level 3)	*"It is not really that bad honestly, I have been through worse situations before."*
Mode 0 (Contractions + Hedging)	*"Honestly, it's not really that bad, I've been through worse situations before."*
Mode 1 (Style Shift + Filler)	*"It is not quite that bad, I have been through worse situations before I guess."*
Mode 2 (Hesitation)	*"It is not really that bad honestly... I have been through worse situations before..."*

Table 9: Example of round-robin lexical mutations for a Disavowal (Level-3) seeker utterance. The core signal (minimization ("not that bad") and historical comparison ("been through worse")) is preserved across all mutations, ensuring label validity.



(a) Before post processing

(b) After post processing

Figure 2: Approximate row-normalised confusion matrices (OOF). Post-processing successfully shifts residual majority-class prediction bias away from L7, noticeably improving minority recall along the diagonal.

Model	BERT-family (Supervised)				Zero-Shot LLM Inference				LLM Fine-Tuning (LoRA)			
	Acc	macro F1	Input / Loss	Ep	Acc	Macro Prec	Macro Recall	macro F1 (1–8)	Acc	Precision	Recall	F1
MentalBERT-base	n/a	0.240 ✓	Flat SEP k=5, Wtd CE, LR 2e-5	5								
MentalRoBERTa-base	n/a	0.2200	Flat SEP k=10, Wtd CE, LR 1e-5	10								
MentalBERT+RoBERTa (ens.)	n/a	0.240 ✓	Flat SEP k=10, Wtd CE, LR 2e/1e-5	5/8								
DeBERTa-v3-base (5-fold)	0.591	0.2358	[CTX]/[TGT], Wtd CE, LR 1.5e-5	8								
RoBERTa-base	0.513	0.2697	[CTX]/[TGT], Wtd CE, LR 2e-5	6								
RoBERTa-base (len=320)	0.479	0.2763	[CTX]/[TGT], Wtd CE, LR 1.5e-5	6								
RoBERTa-base + OS (unif.)	0.618	0.2893	[CTX]/[TGT], Std CE + OS(120), LR 2e-5	6								
RoBERTa-base + OS (tgt.)	0.614	0.2430	[CTX]/[TGT], Std CE + tgt. OS, LR 2e-5	8								
Qwen3-8B					0.3422	0.1773	0.1670	0.1536				
Llama 3.1-8B					0.3036	0.1958	0.1999	0.1584				
Ministral-8B					0.2406	0.2517	0.1481	0.0841				
Mistral-7B-v0.3 (Kaggle)									0.1444	n/a	n/a	0.1023 (0.2054†)
Ministral-8B (Local)									0.6471	n/a	n/a	0.1474
Qwen3-8B ★									**0.6419**	**0.4003**	**0.3958**	**0.3917 ★**

Table 10: Consolidated results across all model families. Empty cells indicate the model was not evaluated under that paradigm. ✓ denotes selected ensemble members. ★ denotes the best result. Rows marked † are external baselines from the task paper (Na et al., 2026a); all others are our own internal tuning experiments.